Pattern Recognition, Volume 46, Issue 9, September 2013, Pages 2420–2436
DOI:10.1016/j.patcog.2013.02.008

A Comparison of 3D Interest Point Descriptors with Application to Airport Baggage Object Detection in Complex CT Imagery

Greg Flitton, Toby P. Breckon, Najla Megherbi

***Abstract*—We present a comparison of 3D feature descriptors with application to threat detection in Computed Tomography (CT) airport baggage imagery. The detectors range in complexity from a basic local density descriptor, through local region histograms and 3D extensions to both to the RIFT descriptor and the seminal SIFT feature descriptor. We show that, in the complex CT imagery domain containing a high degree of noise and imaging artefacts, an object recognition system using simpler descriptors appears to outperform a more complex RIFT/SIFT solution. Recognition rates in excess of 95% are demonstrated with minimal false positive rates for a set of exemplar 3D objects.**

***Index Terms*—CT baggage scan, threat detection, object recognition, 3D feature descriptors, CT object recognition, 3D SIFT**

I. Introduction

X-RAY type technologies have been used for airport security checks for several decades but the use of computer vision within this domain is limited to techniques that purely aid human baggage screeners [1]. Heightened regard to the detection of complex articles within baggage and parcels for air transit and other forms of transportation has led to an increased interest in the use of automatic recognition strategies. Items of interest can be generally difficult to detect within this environment due to a range of orientation, clutter and density confusion in a traditional 2D X-ray projection [2]. An example of this is shown in figure 1 where we see (a) an example bag (photograph), (b) an overhead 2D X-ray revealing an item of interest within and (c) a different scan of the same bag with the item of interest in an orientation that does not reveal its salient features. This potential problem of object self occlusion (figure 1c) is a limitation of 2D X-ray scanners which makes detection (automatically or by human operators) particularly challenging. In this work we specifically look at the use of increasingly popular Computed Tomography (CT) volumetric imagery where a three dimensional voxel image of the baggage/parcel item is obtained in an attempt to overcome some of these issues.

Recent advances in imaging technology now facilitate the use of dual energy CT scanners for the real time scanning of bags in airport baggage/parcel handling operations [3]. It is from these scanners that we obtain a series of image slices through the bag which can be reconstructed as a traditional CT 3D volume akin to those encountered within medical CT imaging [4]. Prior work on the automatic recognition of objects within this complex 3D volumetric imagery is limited [5], [6]. The work of [5] took 3D CT volumes and attempted recognition of an item of interest but reduced the problem to two dimensions by looking at the item characteristic cross section when extracted from the 3D volumetric image (c.f. 2D X-ray views of figure 1). By contrast, [6] explicitly investigated the use of a 3D SIFT descriptor for object recognition with some reasonable results. Here we examine a range of such descriptors and investigate the quality of detection achieveable over a quantifiable larger data set.

(a) Example bag (b) X-ray orientation 1 (c) X-ray orientation 2

Figure 1: Bag and X-rays

It is important at this stage to remember a key aspect of the practicalities of the baggage scanning scenario with relation to the rates of detection: in general we require a high true positive rate (to ensure that true threats are detected) but a low false positive rate (to maximize scanning throughput and additionally minimize impact on the aviation/transport industry).

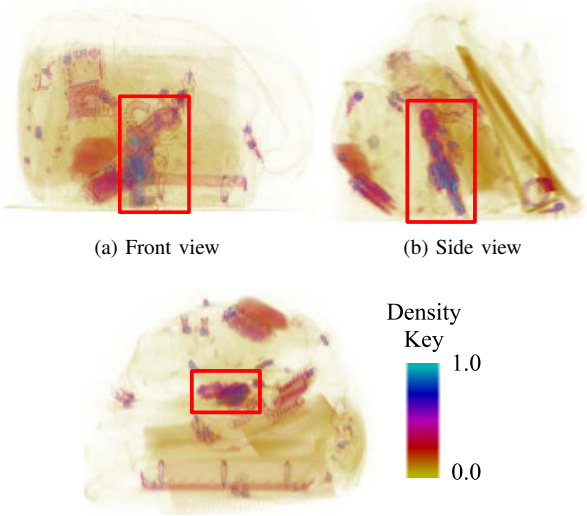
A. Complex CT Volumetric Imagery

An example of a 3D scan of an item of baggage is shown in Figure 2 where we see the presence of an item of interest amongst more general cluttered items. Within figure 2 the data is rescaled to the continuous range $\{0.0 \Rightarrow 1.0\}$ from the original integer CT scanner output (key as shown).

The type of baggage scanner machine used to capture the CT volumetric imagery for this work is primarily aimed at dual energy explosives detection [3]. As a result of this primary (non object recognition based) objective two additional

This project is funded under the Innovative Research Call in Explosives and Weapons Detection (2007), a cross-government programme sponsored by Home Office Scientific Development Branch (HOSDB), Department for Transport (DfT), Centre for the Protection of National Infrastructure (CPNI) and Metropolitan Police Service (MPS). The authors are grateful for additional support from Reveal Imaging Technologies Inc. (USA).

The authors are with the Applied Mathematics and Computing Group, School of Engineering, Cranfield University, Bedfordshire. UK.



(a) Front view (b) Side view

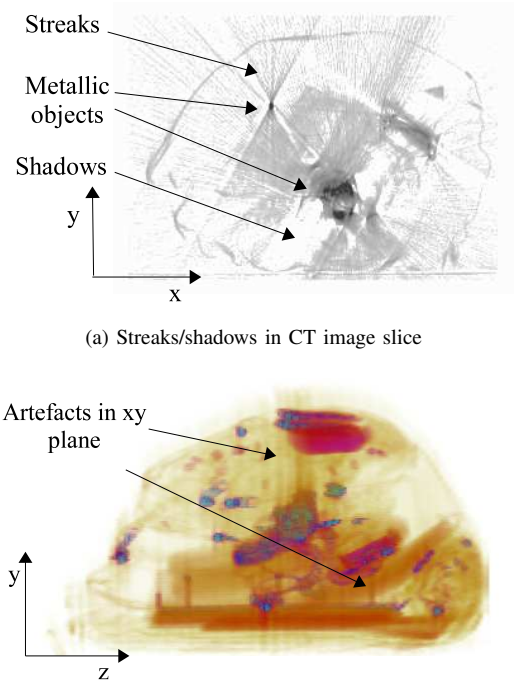
(c) Top view

Figure 2: 3D volume of complex bag containing a revolver

consequences are generally suffered within the imagery :- (1) the presence of metal items causes significant artefacts within the imaging (figure 3) and (2) the resolution is anisotropic and limited to [1.6mm $\times$1.6mm $\times$ 5mm]. The metal artefacts radiate out in the *x-y* plane and do not remain consistent from one scan to another if the metallic region changes orientation. The 5mm resolution in the Z direction will influence the size of target object than can be recognized. Both of these factors are primarily due to the needs of high baggage throughput (speed) and the primary directive of explosives detection such that image quality is sacrificed. It is noted that the artefacts and sampling attributes are significantly different to the current state of the art within medical imaging where the constaints of throughput and the need for dual energy materials detection are not forthright.

Although prior work has looked at the removal of metal artefacts in medical CT imagery [7], [8], [9] this has not been explicitly considered within this work due to constraints on access to the raw CT projection data. Additionally we recognize that the poor resolution gives rise to stair step artefacts [10], [11]. Although this poses significant challenges for recognition we consider here the limitation in resolution to be similar to the scale invariance challenge addressed by various interest point feature descriptors in 2D [12], [13], [14] and additionally the unpredictable nature of the metal artefacts to be akin to that of recognition in the presence of occlusion - an area in which such interest point detectors [12], [13], [14] perform well. Complex imagery of this nature containing dense collections of man made objects scanned at low resolution and in the presence of metal artefacts has not previously been considered within any work on automated 3D recognition.

We choose to resample the anisotropic volumes to create cubic voxels of uniform 2.5mm dimension using cubic spline interpolation. We do not hard limit the interpolation results to the range $\{0.0 \Rightarrow 1.0\}$ and this has the consequence that the working voxel value range is extended to $\{-1.0 \Rightarrow 2.0\}$. Use of this extended voxel value range in subsequent descriptor formulations (Section III) needs to be noted.



(a) Streaks/shadows in CT image slice

(b) Artefacts in xy plane only

Figure 3: Metal artefacts in CT baggage imagery

B. Object Detection in Complex CT Volumetric Imagery

Object detection using interest points and descriptors is a well known approach. The work of [15] proposed the use of Harris features [16] as points of interest in a grayscale image. The interest points were then characterized using a range of rotation invariant descriptors that were then stored in a hash table. Recognition comes by generating the interest points and their descriptors in an image and then looking them up in the hash table. The work of [17] introduced the Scale Invariant Feature Transfrom (SIFT) with the aim of object recognition. Refinements in [12] have led to the SIFT approach being widely used for object recognition in 2D images.

A 3D extension of the SIFT algorithm has been recently presented in the literature by a number of authors [18], [19], [20], [21]. Firstly, Scovanner et al. [18] used a form of 3D SIFT to assist in 3D video volume analysis followed by Cheung and Hamarneh [19] who created a 3D SIFT variant to aid in medical image alignment. Ni et al. [20] also extended SIFT to a 3D formulation, derived from [18], for use in 3D ultrasound panoramic imagery. It is noted that all of these approaches suffer from a fundamental limitation in their consideration of orientation - the definition of orientation in 3D is incorrectly taken as the direction formed by two angles (azimuth, elevation) in [18], [19], [20]. Here, to correctly orientate an object in 3D, we consider three angles - azimuth, elevation and tilt. As shown in figure 9a, three angles are

required to correctly orientate an object. Figure 9b shows an example of this with three pistols aiming in the same direction (given by azimuth and elevation) but with differing orientation (given by the addition of tilt). This prior error of [18], [19], [20] was previously noted by Allaire et al. [21] and corrected: their subsequent results indicated that the additional tilt angle improves matching as expected. Noteably this error originated from the work of [18], [19], [20] as a problem of image registration as opposed to explicit object recognition: a theme also followed by [21]. Here, by contrast to these earlier works, we fully extend SIFT to 3D for the explicit application of object recognition, taking into consideration the full definition of 3D orientation not considered in earlier works [18], [19], [20]. We also compare the system performance using 3D SIFT to that obtained with other descriptors. This extends our previous work of [6].

In this work we explicitly consider the detection of rigid objects within low resolution, noisy, complex volumetric CT imagery and we examine a range of 3D interest point descriptors for this task. This is facilitated by the use of a traditional approach whereby a reference volume object is identified and pose estimated within a given unknown volume. The range of descriptors evaluated for this task range from the use of simple density statistics to full 3D extensions of established interest point descriptors from 2D works [12], [22]. We detail firstly the detection and localization of these descriptors prior to outlining the variants which we go on to present in a range of comparative results.

II. 3D INTEREST POINT DETECTION AND LOCALE

We use interest points and local descriptors as the basis for our object recognition algorithm as these methods have been demonstrated in a variety of fields with high degrees of success [17], [23], [24], [25]. We will now outline our approach for interest point location and local neighbourhood definition.

A. Interest Point Detection

The same method of interest point detection is used for each descriptor being tested so that relative system performance is determined by the choice of descriptor rather than interest point detector. We use a 3D extension to the SIFT algorithm [12], as described in [21], to determine the location of interest points. Given a 3D input volume $I(x, y, z)$ and a 3D Gaussian filter $G(x, y, z, k\sigma)$ we form multi-scale Difference of Gaussian (DoG) volumes as follows:

$$DoG(x, y, z, k) = I(x, y, z) \star G(x, y, z, k\sigma_s) - I(x, y, z) \star G(x, y, z, (k-1)\sigma_s) \quad (1)$$

where k is an integer in the range $\{1..5\}$ representing the scale index, $\sigma_s = \sqrt[3]{2}$ and (x, y, z) are defined in voxel coordinates. Subsequently a three level pyramid $(L = 0, 1, 2)$ is built up by subsampling the Gaussian filtered volume for $k = 4$ and repeating the process.

In a similar vein to the original 2D SIFT methodology [12], DoG local extrema are then located. This requires that a voxel be either a maximum or minimum when compared to its neighbouring voxels. Given that each voxel has a $3 \times 3 \times 3$ local neighbourhood it follows that there are 26 voxels for comparison. It is also a requirement that the voxel is a maxima or minima when compared to the 27 neighbourhood voxels in the scale space DoG volumes both above and below $(k+1,\ k-1)$. The locations of these extrema form a candidate set of interest point locations.

From this candidate set a number of points are rejected for poor contrast if their density is below a threshold, τ_c $(\tau_c = 0.05)$. This removes some erroneous points that are likely to produce unstable descriptors and additionally, in the case of CT volumes, points associated with metal artefacts. A second stage of candidate point rejection also takes place for points which are poorly localized on an edge. These points are likely to produce unstable descriptors in the presence of noise. A 3×3 Hessian matrix describes the local curvature at the candidate point:

$$H = \begin{bmatrix} D_{xx} & D_{yx} & D_{zx} \\ D_{xy} & D_{yy} & D_{zy} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix} \quad (2)$$

where D_{ij} are the second derivatives in the DoG volume. Both [21] and [20] derive a measure to reject points using the Trace and Determinant of H where:

$$Trace(H) = D_{xx} + D_{yy} + D_{zz} \quad (3)$$

$$Det(H) = D_{xx}D_{yy}D_{zz} + 2D_{xy}D_{yz}D_{xz} - D_{xx}D_{yz}^2 - D_{yy}D_{xz}^2 - D_{zz}D_{xy}^2 \quad (4)$$

It can be shown [21], [20] that the following equation can then be used to reject points:

$$Reject\ if \quad \frac{Trace^3(H)}{Det(H)} < \frac{(2\tau_e + 1)^3}{(\tau_e)^2} \quad (5)$$

We use a value of $\tau_e = 40$ and, hence, points where $\frac{Trace^3(H)}{Det(H)} < 332.15$ are rejected.

Finally a subvoxel estimate of the extrema true location is achieved using quadratic interpolation on the DoG volumetric data.

B. Local Point of Interest Neighbourhood Function

Following from the identification of interest point locale we now define a localized neighbourhood function, extending this from earlier work in 2D [12].

A Gaussian window function, $w(d, \sigma)$, is used to limit the contribution of voxels around the point of interest to those in the local neighbourhood:

$$w(d, \sigma) = exp\left[-\left(\frac{d}{\sigma}\right)^2\right] \quad (6)$$

where d is the voxel distance from the point of interest to the contributing voxel and σ is used to determine the extent of the local contribution. The use of this function is given with each of the following descriptor formulations. It should be noted that, given the definition of distance in voxel units, this window will remain consistent with the resolution of the volume being examined.

III. 3D Point of Interest Descriptors

Following interest point detection we now wish to characterize the local neighbourhood. We detail a range of approaches for this characterization in increasing levels of complexity from a simple local density average, density and gradient histograms, leading on to 3D extensions to RIFT [22] and SIFT [12].

A. Simple Density Descriptor, (D)

The density descriptor is a simple Gaussian average around the point of interest as shown in equation 7:

$$D_I = \frac{\sum_k \rho_k . w(d_k, \sigma)}{\sum_k w(d_k, \sigma)} \tag{7}$$

for voxel k, a voxel distance d_k from the interest point location with a density ρ_k. The local neighbourhood function, $w(d_k, \sigma)$, is as defined in II-B.

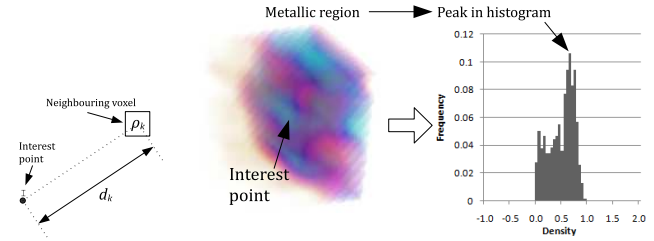
This is a simple detector and is included for comparison to its more complex counterparts.

B. Density Histogram Descriptor, (DH)

By contrast this second descriptor defines the local density variation at a given interest point as an N bin histogram defined over a continuous density range. The density range is $\{-1.0 \Rightarrow 2.0\}$, in line with the resampled cubic voxel volume, and is split into N_{DH} bins resulting in each bin having a width of $3.0/N_{DH}$. The voxel density for point k is ρ_k and this is used to determine which histogram bin is active. Given the local area function $w(d_k, \sigma)$, defined in Section II-B, an addition of $w(d_k, \sigma)$ is made to the appropriate histogram bin where d_k is the voxel distance from the point of interest to voxel k. The descriptor is normalized to unity area on completion. Figure 4a shows an example point of interest, I, with one of its neighbouring voxels of density ρ_k. Figure 4b shows an example of a density histogram derived from an interest point that is located near a metallic region. It can be seen from this that the resulting density histogram has a peak due to the high concentration of metal within the neighbourhood.

C. Density Gradient Magnitude Histogram Descriptor, (DGH)

In a variant of the previous descriptor, here we calculate the density *gradient* magnitude in the neighbourhood of the interest point and then accumulate these in a histogram. The



(a) Local Neighbourhood Voxels (b) Example Density Histogram Descriptor

Figure 4: Density Histogram Calculation

density gradient magnitude is calculated for all voxels in the volume using a central difference formulation to ensure that the gradient location is aligned with the voxel grid. The density gradient magnitude range is $\{0.0 \Rightarrow 4.0\}$ (given a voxel dimension of 0.25cm, the vast majority of gradient values lie below a value of 4.0) and is divided into N_{DGH} bins resulting in each bin having a width of $4.0/N_{DGH}$. The voxel gradient magnitude for voxel k is δ_k and this is used to determine which histogram bin is active. Once the active histogram bin is determined an addition of $w(d_k, \sigma)$ is made to the corresponding histogram entry, with $w(d_k, \sigma)$ again defined in Section II-B. The descriptor is normalized to unity area on completion as per the previous descriptor (Section III-B). It is notable here that, due to the rotational variance of the objects under consideration for detection, the gradient *magnitude* is used rather than the gradient *orientation* approach frequently used for recognition tasks in 2D [26].

Figure 5 shows the same point of interest as for figure 4b but now with the density gradient histogram being formed. It is not as obvious how the histogram relates to the volume given the noisy conditions of the imagery.

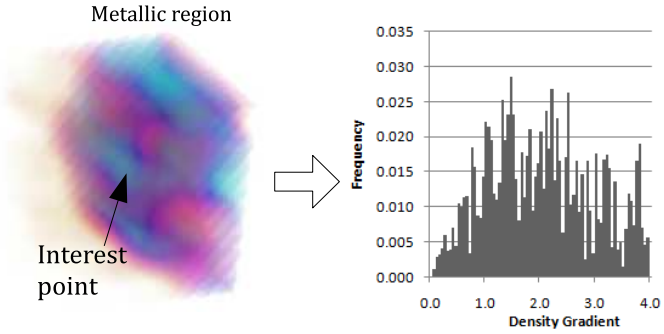


Figure 5: Density Gradient Histogram Calculation

D. Rotation Invariant Feature Transform, (RIFT)

The work of [22] developed the Rotation Invariant Feature Transform (RIFT). The RIFT descriptor examines the local neighbourhood gradients with reference to a radial vector emanating from the point of interest. Histograms are constructed from the gradient orientation and magnitude. Multiple histograms are derived following segmentation of the local

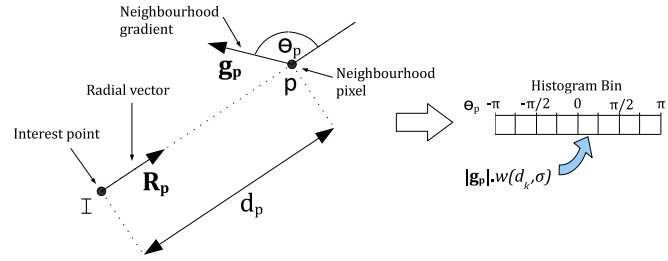
neighbourhood into a series of rings centred on the point of interest. RIFT has been shown to operate well in standard 2D imagery and is used in our work as it is more complex than the simple histograms described above, but is not as complex as the SIFT descriptor [22], [12].

Before describing our extension of RIFT to 3D we consider our variant concretely in 2D.

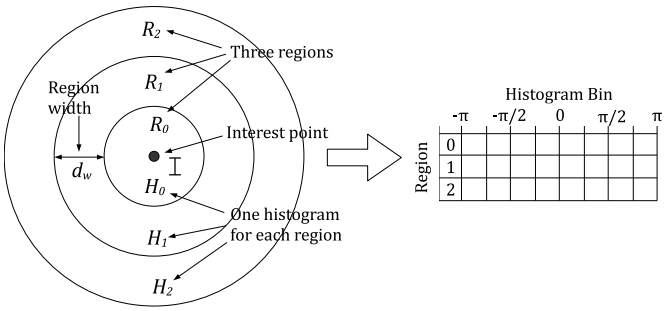
Figure 6a shows a point of interest, I, and neighbouring region. For each neighbouring pixel, p, a unit vector in the direction $\overrightarrow{Ip}$ is calculated: $\mathbf{R_p}$. The gradient at pixel p is $\mathbf{g}_p$. The angle between the gradient $(\mathbf{g}_p)$ and radial vector $(\mathbf{R_p})$ is θ_p. A histogram is constructed based on values of θ_p in the range $[-\pi : \pi]$. There are N_b bins in this histogram representing angular regions $2\pi/N_b$ radians in size. For each gradient and angle an addition to the histogram of $|\mathbf{g}_p| .w(d_k, \sigma)$ is made as shown in figure 6a. Note again that the function $w(d_k, \sigma)$ limits the contribution to the local neighbourhood. In addition to the histogram, N_r rings, of width d_w pixels, are also defined as shown in figure 6b (with $N_r = 3$). One histogram is generated for each region and each histogram is normalized by the area of its ring to prevent bias to regions of greater area. The complete descriptor is normalized to unity. The resultant descriptor has $N_r \times N_b$ elements.

The extension of this descriptor to 3D is straight forward noting that, due to rotation symmetry in 3D, the radial histograms only cover values of θ_p in the range $[0 : \pi]$ and the normalizations refer to region *volumes* rather than areas. One additional normalization is required in the move to 3D: the histogram summations are normalized by bin surface area to remove bias towards equatorial bins. Figure 7 shows an example with 4 bins per histogram: Bins A, B, C and D. If the volume has unit radius, bins A and D have a surface area of $\pi\left(2-\sqrt{2}\right)$, whereas bins B and C have an area of $\pi\sqrt{2}$. These areas are used to normalize the summations for each bin. This step is not required in the 2D case as all histogram bins have the same sector angle.

As with other descriptors, the final step is to normalize the complete descriptor to unity.



(a) 2D Radial Geometry



(b) 2D Radial Regions

Figure 6: 2D RIFT Descriptor

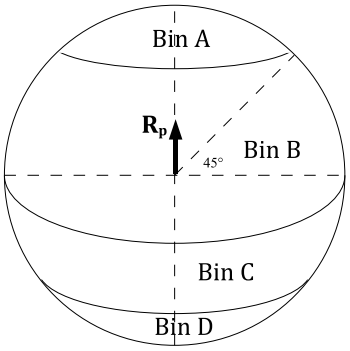


Figure 7: 3D RIFT bin normalization

Figure 8 shows the RIFT descriptor generated for the same metallic region as used in the Density Histogram and Density Gradient Histogram explanations (figures 4b/5). This plot shows that, for this example, the gradients tend to point toward to point of interest rather than away.

E. 3D SIFT

This descriptor is closely modelled on that used in [21], [6]. We briefly outline our 3D SIFT extension detailing the keypoint orientation and description based upon the initial interest point detection steps and local neighbourhood function as outlined in Section II. Here as an extension to previous work on 3D SIFT [18], [19], [20], we follow the work of [6] and fully consider object recognition taking into consideration 3D orientation in terms of azimuth, elevation and tilt as illustrated in figure 9.

1) Keypoint Orientation: Once a keypoint location is determined (Section II) the volume gradients are examined in a two stage process to locally establish an invariant orientation in the subsequent description. A *direction* in 3D space is defined by the azimuth and elevation angles whereas an *orientation* is defined by the addition of a third angle: tilt (see Figure 9).

The first step is to determine the dominant *direction* for the keypoint. A 2D histogram is produced by grouping the Gaussian filtered volume gradients in bins which divide azimuth and elevation into 45° sections, as shown in Figure 10a (sphere) and Figure 10b (resulting 2D histogram bins). Consequently there are N_a $(N_a = 8)$ azimuth bins and N_e $(N_e = 4)$ elevation bins. A regional weighting is applied to the gradients according to their voxel distance from the keypoint location: we apply a Gaussian weighting of $exp\left[-\left(2r/R_{max}\right)^2\right]$ for voxels a distance r from the keypoint location. Points further than R_{max} voxels from the location are ignored in the current formulation. From a geodetic viewpoint (Figure 10a) it can

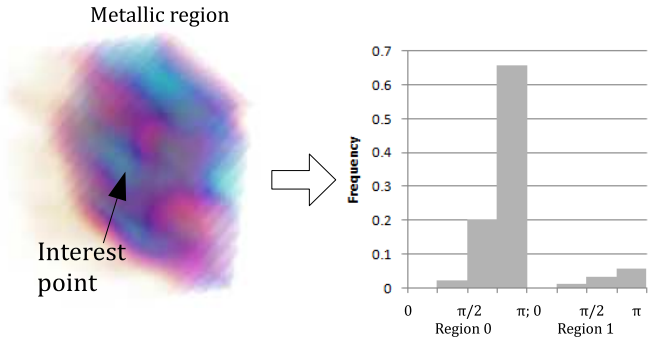
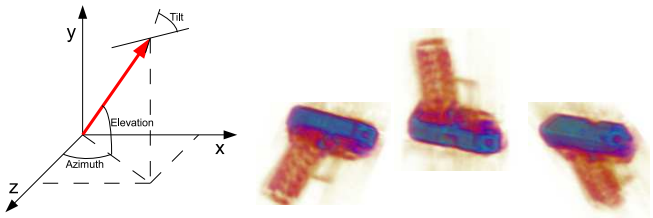
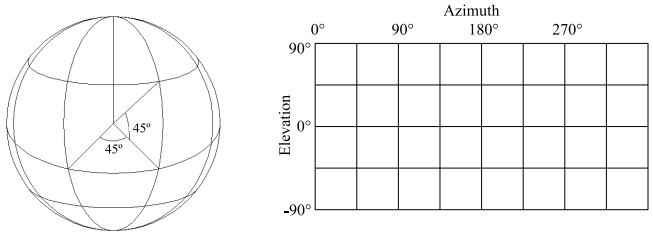


Figure 8: RIFT descriptor example



(a) Our definitions (b) Pistols pointing in same direction but with differing orientation (ie tilt)

Figure 9: 3D Orientation requires three angles: Azimuth, Elevation and Tilt



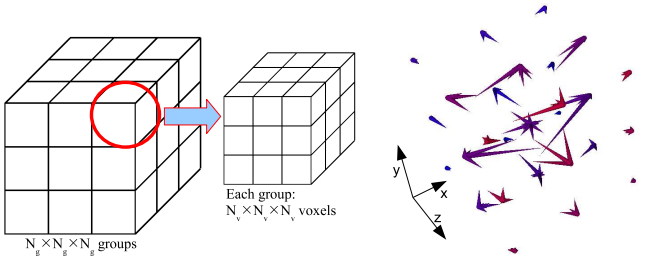
(a) splitting Azimuth/Elevation into 45° bins (b) resultant 2D histogram bins

Figure 10: Direction Histogram

be seen that bins near the *equator* in this formulation are larger than those at the poles and this will bias the resulting histogram. This bias is compensated for by normalizing each histogram bin by its solid angle [18]. The output histogram is then smoothed using a Gaussian filter to limit the effects of noise and the dominant directions are determined by searching for peaks and are refined using interpolation. Peaks in this 2D histogram within 80% of the largest peak are also retained as possible secondary directions in line with the formulation of [12].

The second step is to determine the *orientation* by calculating the tilt angle for each derived direction. This is achieved by re-orientating the volume around the keypoint and calculating a 1D histogram that resolves the gradients orthogonal to the dominant direction. This histogram is again built in 45° bins using the same regional weighting method as for the direction histogram. Peaks in the tilt histogram are used, with interpolation, the derive an estimate of keypoint tilt. Again, peaks within 80% of the largest peak are retained to give secondary orientations. Overall, in this formulation, we see that keypoints may have more than one possible orientation that will require description.

2) Keypoint Description: Once the orientation has been determined the point of interest can be described. In our case we build a $N_g \times N_g \times N_g$ grid of gradient histograms, with each histogram being computed from a $N_v \times N_v \times N_v$ voxel grouping as shown in figure 11a. Each gradient histogram is derived by splitting both azimuth and elevation into 45° bins, as described in Section III-E1. Consequently, each descriptor, normalized to unity, contains $N_g^3 \times N_a \times N_e$ elements. The final visualization of such a descriptor is shown in figure 11b as a 3D grid of gradient histograms.



(a) Voxel groupings for descriptor (b) $3 \times 3 \times 3$ 3D SIFT descriptor

Figure 11: 3D SIFT Descriptor Formulation

IV. Object Detection Methodology

An overview of descriptor generation is shown in figure 12 where we see the separation of interest point detection from descriptor generation which, in our comparison for this work, can be performed in a number of different ways (as described in Section III). Interest point locations for an input volume are generated using the SIFT derived methodology described in Section II. Descriptors for each volume are generated using these locations. The location of the keypoint is stored as part of the descriptor to facilitate a relative position consistency check in the recognition methodology.

An object detection system methodology is shown in figure 13. Here we start with a known reference item from which descriptors are calculated. A candidate baggage item is received and processed to determine its descriptors. The matches between the reference and candidate item are filtered in an attempt to retain true matches and remove false matches. The output set of matches from this process are referred to as the correspondence set.

Two methods are used when forming the correspondence set:

a) The method of [12] where a match is accepted to the correspondence set if the ratio of the first and second best match distances is less than 0.8. We refer to this method as the distinction method. We consider this process from the

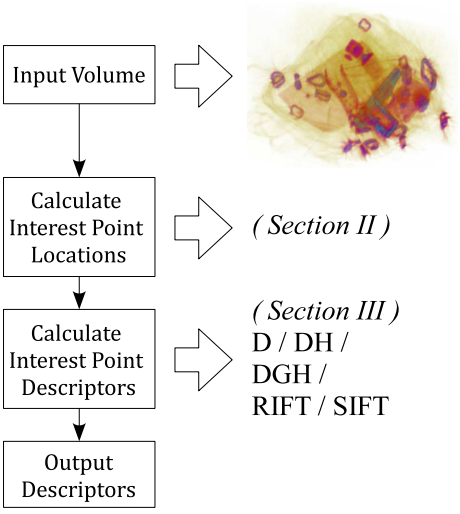


Figure 12: Descriptor Generation

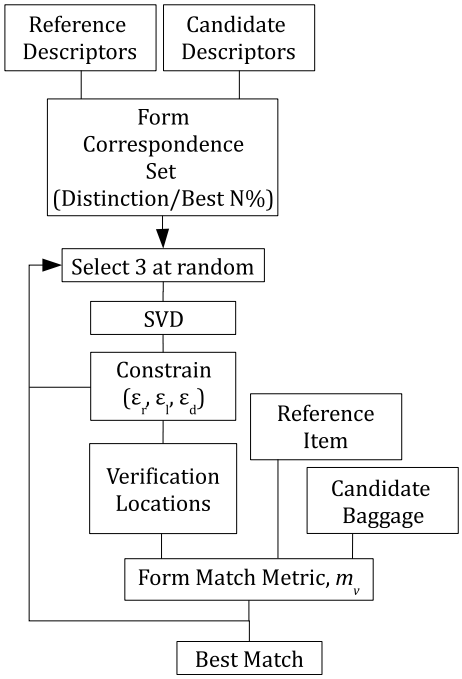


Figure 13: Object Recognition Methodology

candidate to the reference. i.e. a candidate/reference pair is added to the correspondence set if it is distinct compared to matches between the same candidate and the other reference descriptors.

b) We reorder the matches in order from lowest Euclidean distance upwards. We then choose a fixed percentage of the best matches as the correspondence set. We refer to this method as the percentile method, with parameter p defining the percentage of matches used.

Given the large number of possible false matches in this formulation we make use of RANSAC (RANdom SAmpling and Consensus) [27], to find an optimal match using the correspondence set as the input. The RANSAC algorithm [27] has been shown to cope well in the presence of significant outliers (here highly prevalent due to noise). This RANSAC formulation is used to select a set of three possible matches from the correspondence set from which a 3D transformation is derived using a common Singular Value Decomposition (SVD) approach [28].

Following estimation of the transformation we check to see if the three RANSAC selected matches are consistent in a number of ways:

a) the reference set and candidate set should be similar shapes: relative distance errors should be less than ϵ_r $(\epsilon_r = 10mm)$
b) the reference set and candidate set should be in similar locations: absolute location errors should be less than ϵ_l $(\epsilon_l = 10mm)$
c) the reference set and candidate set should have similar densities: density errors should be less than ϵ_d $(\epsilon_d = 0.1)$

It should be noted that the one to one relationship between voxel measurements and real world distances allows the tolerances ϵ_r and ϵ_l to be specified in real world measurements (i.e. mm). These constraints aid the matching process by quickly rejecting poor quality selections prior to the verification stage.

If the relative distance and density criterion is passed a secondary verification is performed. Locations in the reference object with a density above a threshold τ_d $(\tau_d = 0.15)$ are recorded to form a set of density verification locations. The threshold is applied in order to reduce the number of low density artefacts in the verification stage. The verification locations are transformed into the candidate baggage item space using the transform estimate provided by the SVD formulation. Given N_v verification points we then form a quality of match metric, m_v, by examining the density differences between the verification locations in the reference item and the candidate baggage item:

$$m_v = \frac{\sum_{k=1}^{N_v} |\rho_k - \psi_k|}{\sum_{k=1}^{N_v} \psi_k}$$

where ψ_k is the density at the k^{th} verification point in the *reference* item and ρ_k is the density of the voxel closest to the k^{th} transformed verification point in the *candidate* baggage item. The measure is normalized by the sum density of the verification points in the reference item, as shown, to provide a metric that does not vary too greatly between different reference items.

The set of descriptors for comparison, described in Section II, were computed using the parameter settings shown in Table I. The results of this comparison using the proposed object detection methodology and the parameter settings listed in Table I are presented in the next section.

Table I: Descriptor Settings

Descriptor	Settings	Dimension
Density	$\sigma = 1.0$	1
Density Histogram	$\sigma = 3.0$, $N_{dh} = 60$	60
Density Gradient Magnitude Histogram	$\sigma = 3.0$, $N_{gh} = 80$	80
RIFT	$\sigma = 3.0$, $N_b = 4$, $N_r = 2$, $d_w = 3.0$	8
SIFT	$N_g = 3$, $N_v = 3$, $N_a = 8$, $N_e = 4$	864

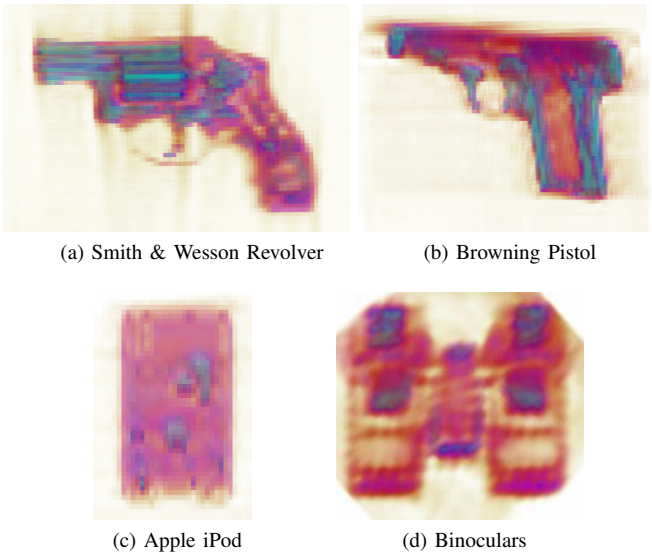
Table II: Items scanned

Baggage Item Contents	Scans in Collection
Smith & Wesson Revolver + Clutter	21
Browning Pistol + Clutter	30
Apple iPod + Clutter	15
Compact Binoculars + Clutter	14
Clutter only	180

V. Results

First we consider the distinction method when forming the correspondence set which is true to [12] rather than the percentile approach that was successfully used in [6].

For this comparative study four target items of interest were used (Smith & Wesson revolver; Browning pistol; Apple iPod; compact binoculars) of which scans are shown in figure 14. Furthermore a mix of baggage types (e.g. holdalls, suitcases, handbags, etc.) containing a variety of clutter items as would be found in a typical airport scenario, including and excluding these items of interest, were scanned using a Reveal Imaging Technologies 3D CT80 scanner. Table II shows the number of baggage items scanned which contained one of these target items or which were left clear of the named targets but still contained regular background clutter.



(a) Smith & Wesson Revolver (b) Browning Pistol

(c) Apple iPod (d) Binoculars

Figure 14: Reference CT Object Volumes Used For Detection

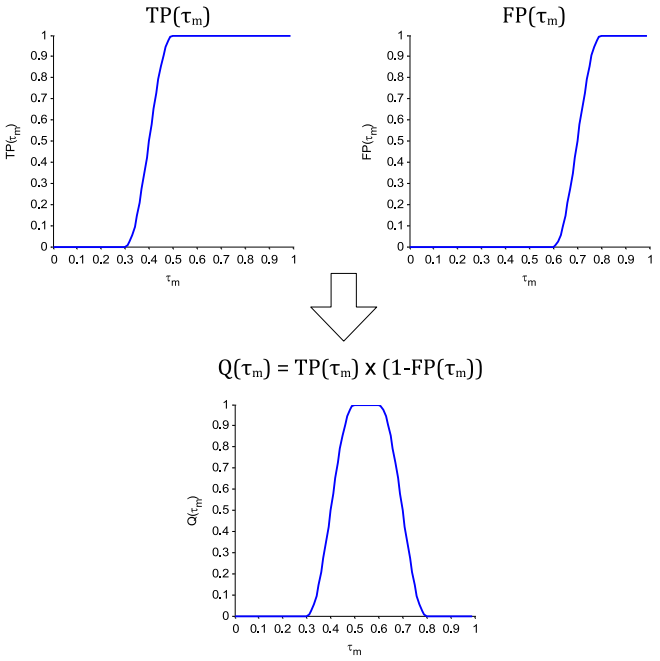
Each baggage item is "searched" using the object detection methodology outlined in Section IV for each of the four reference CT object volumes shown in figure 14. From this each baggage item produces a verification match metric result, m_v, as described in Section IV (a measure of similarity between the reference item and the baggage item). A decision on whether a target item has been detected is made by comparing the verification match metric result, m_v, against a detection threshold, τ_m. Given ground truth knowledge of which baggage items contain the target items and which do not we can calculate both a True Positive detection rate, $TP(\tau_m)$, and a False Positive detection rate, $FP(\tau_m)$, for a given setting of τ_m. Our analysis uses Receiver Operating Characteristic (ROC) plots [29] to investigate the overall system performance as each descriptor type is used. These plots show $TP(\tau_m)$ against $FP(\tau_m)$ and indicate the trade off between true detection of threat items versus false detection as the detection threshold, τ_m, is varied. When producing a numerical performance result we choose to quote the True Positive rate for minimal False Positive rate (<1%) rather than the ROC equal error rate [30] as we feel that this is more applicable to the operating conditions of such a system in an operational security environment (even a moderate False Positive rate is not desirable).

The ROC plot gives one aspect of performance. We also form a plot that shows a measure of tolerance to error in the value of the detection threshold, τ_m, should a fixed value be chosen to decide the presence of the target item. We refer to this as the Threshold Quality, $Q(\tau_m)$, where:

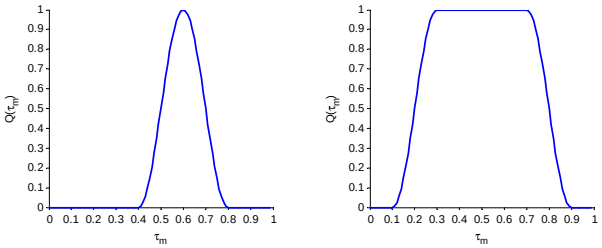
$$Q(\tau_m) = TP(\tau_m) \times [1 - FP(\tau_m)] \quad (8)$$

Figure 15a shows how the True Positive and False Positive rates are combined to form the Threshold Quality. The width of the Threshold Quality plot indicates the separation between the rise in True Positive rate and the rise in False Positive rate. The height of the Threshold Quality peak is also indicative of performance. If the True Positive and False Positive rates are well separated then the Threshold Quality will reach a peak value of 1.0 which would indicate a perfect ROC plot. However, if the True Positive and False Positive transition regions overlap the Threshold Quality peak will be less than 1.0. Figures 15b and 15c show Threshold Quality plots for two systems, both with perfect ROC plots. It can be seen in figure 15b that the Threshold Quality peak is narrow indicating that the True Positive transition region is close to the False Positive transition region. A better scenario is shown in figure 15c where the Threshold Quality peak is broad indicating a large separation between the True Positive and False Positive transition regions. This broad peak indicates

that, when allocating a value to the detection threshold (τ_m), a greater tolerance to error in its assignment exists.



(a) Threshold Quality Derivation



(b) Poor Threshold Quality (c) Good Threshold Quality

Figure 15: Threshold Quality

We now present our results as ROC plots using the legend given in Table III.

Table III: Plot Legend

Descriptor	Legend
Scale Invariant Feature Transform	SIFT
Density	D
Density Histogram	DH
Density Gradient Histogram	DGH
Rotation Invariant Feature Transform	RIFT

First we examine detection performance using the distinction approach of [12] to form the correspondence set and then we will look at the performance using the percentile method proposed in our earlier work of [6].

ROC plots for detection of the revolver, pistol, iPod and binoculars when using the distinction method are shown in figure 16. It can be seen that there is a considerable variation in detection performance between the descriptor types, as well as differing levels of detection of each target item.

For the revolver (figure 16a) the best result using the distinction method is obtained using the RIFT descriptor with a detection rate of ~95% with detection using Density, Density Histogram and Density Gradient Histogram at ~60/70% . The performance of SIFT is poor with a detection rate of ~20%.

The pistol performance is poorer (figure 16b) with a detection rate of ~55% with a negligible false positive rate using the Density Gradient Histogram. This is closely followed by the Density and Density Histogram descriptors (~50%) with RIFT and SIFT both poor (~20%).

The iPod performance is worst (figure 16c) with a detection rate of ~20% using the RIFT descriptor, closely followed by Density, Density Histogram and Density Gradient Histogram (~15%) with SIFT again the worst performing (~5%).

Detection of the binoculars is ~80% (figure 16d) with negligible false positives using the Density Gradient Histogram. Detection using RIFT, Density and Density Histogram descriptors is ~50% with SIFT again worst with a detection rate of ~20%.

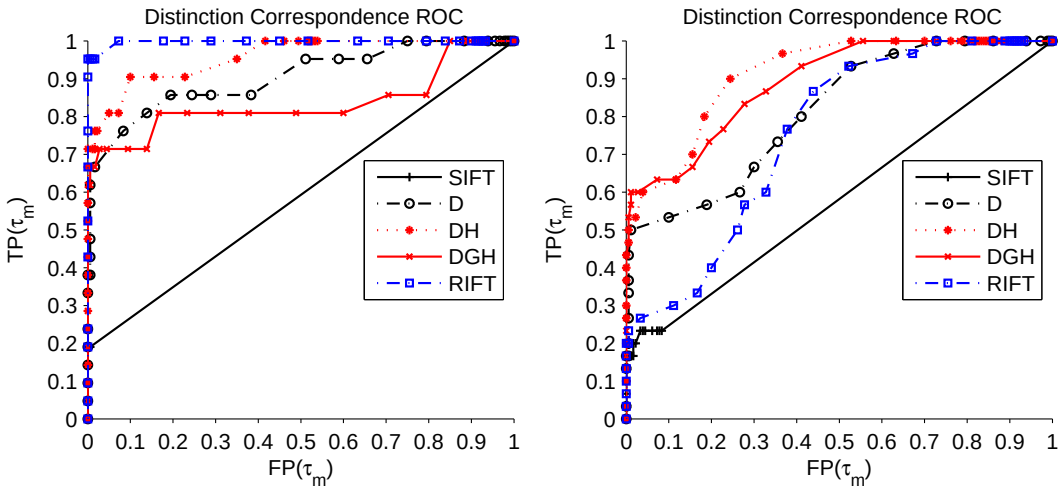
Two immediate questions arise from further consideration of these results:

a) why is the pistol detection rate (~55%) poorer than the revolver (~95%) given that they are similar items in both size and density characteristic?

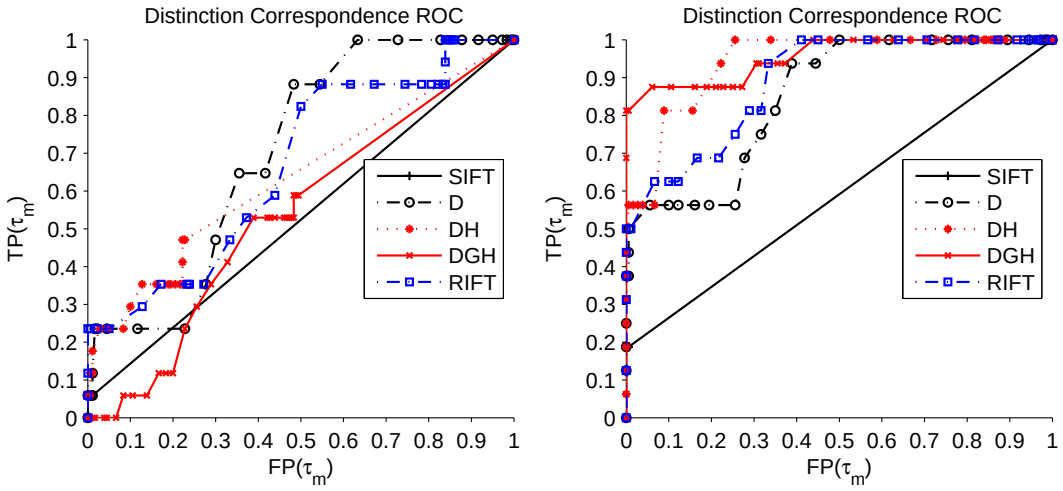
b) why does the use of the SIFT descriptor yield much poorer results when compared to simpler descriptor types?

An investigation into the poor quality of the pistol results compared to those of the revolver indicated that the scan quality of the reference item affects performance. Figure 17a shows the reference used to create the results in figure 16b. Figure 17b shows a different scan of the Browning pistol which is used as a reference. Note in this secondary example (figure 17b) the clarity of the pistol muzzle (A) compared to figure 17a. Also note apparent density differences in the barrel (B), trigger guard (C) and grip (D) caused by metal artefacts and anisotropic scanning. These differences will affect the resulting descriptors, both in value and location, and this has obvious implications for location of similar points in randomly scanned baggage items. The difference between these scans is the orientation of the pistol relative to the CT scanner Z axis, as shown in figure 18. The original pistol reference (figure 18a) was orientated such that the barrel was parallel to the XY plane resulting in the barrel being scanned with a 5mm resolution (the CT slice spacing - see Section I-A). The alternate pistol reference (figure 18b) was scanned such that the barrel was orthogonal to the XY plane resulting in a barrel cross section resolution of ~1.6mm (the slice pixel resolution - see Section I-A).

Figure 19 shows the ROC plot using match distinction to form the correspondence set when using the *alternate* pistol reference. Here we can see a better detection rate of ~85%

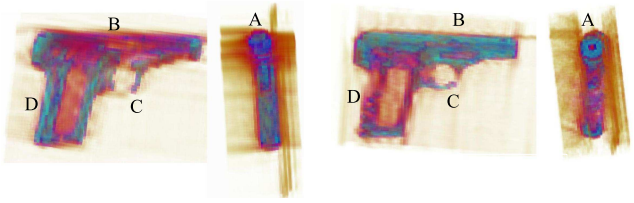


(a) Revolver (b) Pistol



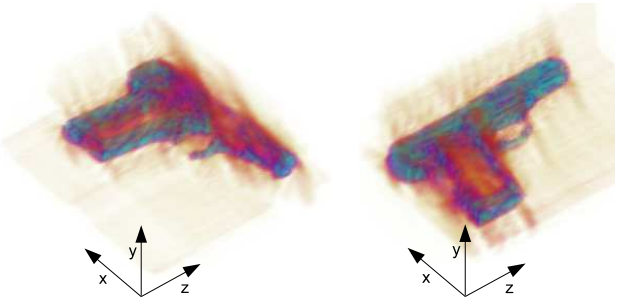
(c) Apple iPod (d) Binoculars

Figure 16: Target item ROC curves using distinction to form correspondence set



(a) First Reference (b) Second Reference

Figure 17: Browning Pistol Reference Item Quality



(a) First Reference Item (b) Second Reference Item

Figure 18: Browning Reference Item Orientation in CT Baggage Scanner

using the Density Histogram descriptor (up from ~50%). The RIFT descriptor has a detection rate of ~70% (up from ~20%) with Density Gradient Histogram at ~60% (from ~55%), Density at ~50% (unchanged) and SIFT at ~20% (unchanged).

We combined the results for both pistol references by choosing the result with the lowest verification match metric value, m_v, to observe if the combination would provide increased levels of performance. Figure 20 shows the ROC plot for this situation where we can see that an improvement does occur (compared to the individual reference item results shown in

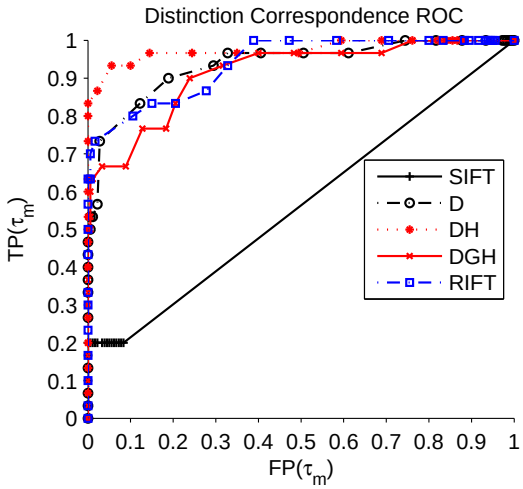


Figure 19: ROC using second Browning pistol as reference

figure 16b and figure 19). The best performance again comes from the Density Histogram with a detection rate of ~90% with negligible false positives (up from ~85%). The performance using the other descriptors is also improved: Density ~75% (up from ~50%); Density Gradient Histogram at ~80% (up from ~60%); RIFT up slightly at ~75% (from ~70%); SIFT at ~35% (up from ~20%).

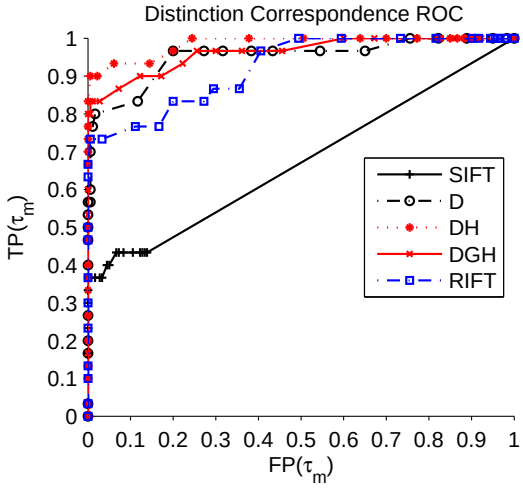


Figure 20: ROC for combination of pistol results

An investigation into why the use of the SIFT descriptor yielded poor detection results was carried out. Analysis of the correspondence set showed that, when using match distinction, very few of the SIFT matches are deemed to be suitable. Table IV shows the mean correspondence set size (as a % of total matches) for each target item and each descriptor when analyzed over the data sets given in Table II. For the Density, Density Histogram, Density Gradient Histogram and RIFT descriptors we see correspondence set sizes between 0.80% and 3.08% of the total number of matches. When compared to these descriptors, the SIFT descriptor has very few matches in the correspondence set: between 0.01% and 0.07%. This is indicative of poor quality descriptors (very few pass the distinction criterion) and it would appear that this restricts its performance: true matches are rejected from the correspondence set and not enough are made available to the object detection method for reliable recognition of the target items.

It is noteable that the use of the distinction method differs from the selection method used in our prior work [6] where significantly improved SIFT 3D object detection results were obtained.

In light of these results and with the support of the earlier work [6] we vary the method used to form the correspondence set away from the seminal 2D SIFT variation [12] and use our alternative percentile method as previously discussed in Section IV. Rather than using the match distinction method we instead sort the matches by match distance and then choose a fixed percentage of the best matches as per [6].

Figure 21 shows the results when the best 2% of matches are chosen to form the correspondence set.

For the revolver (figure 21a) we can see near 100% detection with minimal false positives using Density Histogram, Density Gradient Histogram and RIFT descriptors. Both Density and SIFT descriptors have detection rates ~85%.

Using the second pistol reference (figure 21b) we again see near 100% detection using the RIFT descriptor, closely followed by Density Histogram and Density Gradient Histogram (~90%) with SIFT at ~65% and Density at ~35%.

The iPod detection is still poor (figure 21c), though slightly improved, at ~30% (increased from ~20%) using the Density Histogram, followed by Density Gradient Histogram, RIFT and SIFT at ~20%. The Density descriptor has a detection rate of ~0% using our negligible false positives detection rate definition.

The binoculars show near 100% detection (figure 21d) using RIFT, Density Gradient Histogram and SIFT, with Density Histogram close behind at ~95%. The Density descriptor is again poor with a detection rate of ~0%.

Given a number of ROC plots that appear to show 100% detection rates, mainly due to the limited amount of target items, we can also investigate performance using the Threshold Quality, $Q(\tau_m)$, as the detection threshold, τ_m, is varied (equation 8). Threshold Quality plots relating the the ROC plots in figure 21 are given in figure 22.

Figure 22a shows the plot in the case of the revolver where we see the superior performance of the Density Histogram descriptor and RIFT descriptor over the Density Gradient Histogram descriptor that it is not possible to see in the ROC plots (figure 21a). Both the Density Histogram and RIFT descriptor reach a peak when $\tau_m \simeq 0.45$ and then fall off when $\tau_m \simeq 0.6$. The Density Gradient Histogram only reaches a peak for $\tau_m \simeq 0.55$ and then almost immediately starts to fall away. The implication for this, in a noisy environment,

Table IV: Mean Correspondence Set Size (as % of total matches)

Descriptor	Revolver	Pistol	iPod	Binoculars
Density	2.31 ± 0.10	2.47 ± 0.07	3.08 ± 0.10	1.71 ± 0.11
Density Histogram	0.80 ± 0.31	1.32 ± 0.30	1.20 ± 0.50	0.96 ± 0.27
Density Gradient Histogram	1.55 ± 0.23	1.18 ± 0.23	0.93 ± 0.20	0.81 ± 0.18
RIFT	1.39 ± 0.14	1.05 ± 0.15	1.17 ± 0.20	1.15 ± 0.11
SIFT	0.02 ± 0.01	0.07 ± 0.06	0.02 ± 0.01	0.01 ± 0.01

would be that the Density Histogram and RIFT descriptors would be more reliable than the Density Gradient Histogram.

Figure 22b shows the Threshold Quality for the second pistol reference. Here we see that, although both the RIFT and Density Gradient Histogram descriptors reach a peak of 1.0, they quickly fall away. This does not appear to be as good as the revolver.

Figure 22c shows the results for the Apple iPod. Here we see poor results already indicated by the ROC plot (figure 21c).

Figure 22d shows the results for the binoculars. Here we can see that the RIFT descriptor has the broadest peak, closely followed by the Density Gradient Histogram. The SIFT descriptor, though apparently with near perfect ROC, only just reaches a peak of 1.0 before falling away. The Density Histogram, though apparently not as good in the ROC plot (figure 21d), has the widest peak which would indicate it is more tolerant to detection threshold selection error.

Varying threshold quality gives us an alternative statistical visualization of the relative performance of the different 3D interest point descriptors within this context.

VI. Conclusions

Our results have shown that creation of the correspondence set using the distinction method of [12] is not the best approach in the case of complex CT imagery containing a large number of artefacts. Better results are obtained if the correspondence set is determined by sorting the matches by Euclidean match distance and then taking a fixed percentage of the best matches [6].

Detection of the revolver, pistol and binoculars was achieved with near perfect results although this is more an indication that the number of target items needs to be increased to correctly estimate margins of error in detection. Due to the practical difficulties in obtaining large data sets of the nature considered in this work an extended study over such large data sets is left as an area for future work.

We have shown that an anisotropic scanning system will affect the recognition results. The Browning pistol was scanned in orthogonal orientations and produced very different recognition results. Care thus needs to be taken when choosing a reference item or, as we have demonstrated, multiple reference volumes can be used to improve detection results. The use of multiple reference object scans and methods of determining reference scan quality is also left as an area for future work.

By contrast to the complexity of the 3D SIFT implementation, a simple histogram of density data in the local region of a point of interest provided very good comparative results.

The 3D RIFT descriptor produced good results using the distinction approach to produce the correspondence set and also performed well in the fixed percentage approach. The 3D RIFT descriptor is very concise: only 8 values are stored compared to 864 for 3D SIFT.

The 3D SIFT descriptor can produce good results but it would appear that simpler descriptors (Density Histogram, 3D RIFT) produce better results with the advantage of reduced complexity. It would appear that the 3D SIFT descriptor is not robust in the presence of a large amount of CT artefacts and this is understandable as the artefacts will greatly affect the density gradients upon which dominant orientation is decided and subsequent descriptor histograms are built.

Detection of the iPod was poor. The best result of 30% was achieved using the percentile method ($p = 2.0\%$). It is believed this is due to its lower density which is more easily corrupted by metal artefacts in the baggage item. It is also a fact that the iPod dimensions $(104mm \times 62mm \times 11mm)$ ensure that most descriptors include areas outside the device in their formulation and, as such, are prone to adjacent baggage items influencing the descriptor.

Overall we have shown a comparison of differing 3D point descriptors applied to the problem of object detection in complex 3D CT volumetric imagery. It has been shown that approaches based on simpler density information outperform more complex 3D extensions of common and established point descriptors adapted from 2D image recognition [12], [22]. Further work will investigate the use of multiple objects as a derivative for the reference volume and also the evaluation of quality and artefact information within the imagery.

References

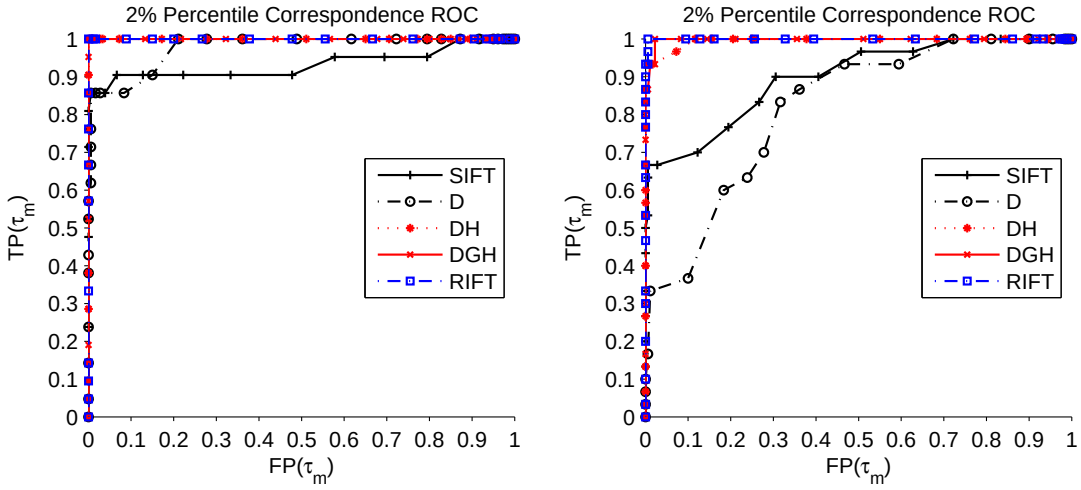
[1] B. Abidi, Y. Zheng, A. Gribok, and M. Abidi, "Improving weapon detection in single energy X-ray images through pseudocoloring," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 36, no. 6, pp. 784–796, 2006.

[2] A. Schwaninger, A. Bolfing, T. Halbherr, S. Helman, A. Belyavin, and L. Hay, "The Impact of Image Based Factors and Training on Threat Detection Performance in X-ray Screening," in *Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008*, pp. 317–324, 2008.

[3] S. Singh and M. Singh, "Explosives detection systems (EDS) for aviation security," *Signal Processing*, vol. 83, no. 1, pp. 31–55, 2003.

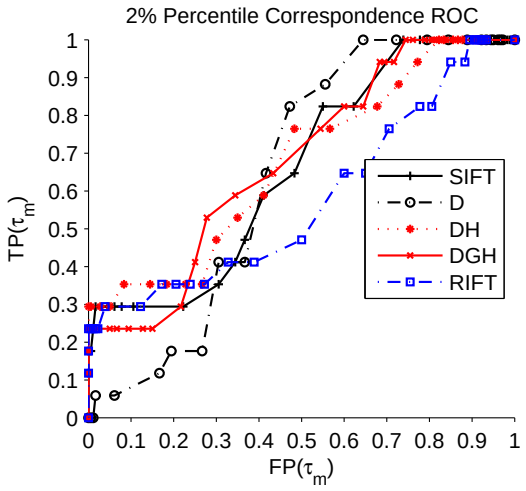
[4] G. Herman, *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*. Springer Verlag, 2nd ed., 2009.

[5] W. Bi, Z. Chen, L. Zhang, and Y. Xing, "A volumetric object detection framework with dual-energy CT," in *IEEE Nuclear Science Symposium Conference Record, 2008.*, pp. 1289–1291, October 2008.

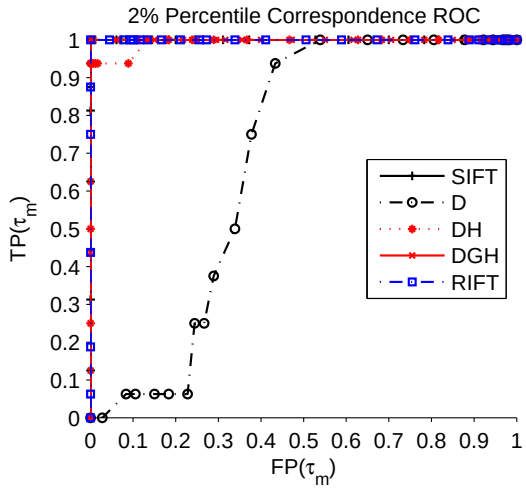


(a) Revolver

(b) Pistol (second reference)



(c) Apple iPod



(d) Binoculars

Figure 21: ROC curves when using percentile matches $(p = 2\%)$ for correspondence set

[6] G. Flitton, T. Breckon, and N. Megherbi, "Object Recognition using 3D SIFT in Complex CT Volumes," in *Proceedings of the British Machine Vision Conference* (F. Labrosse, R. Zwiggelaar, Y. Liu, and B. Tiddeman, eds.), pp. 11.1–11.12, BMVA Press, 2010.

[7] W. Kalender, R. Hebel, and J. Ebersberger, "Reduction of CT artifacts caused by metallic implants," *Radiology*, vol. 164, no. 2, p. 576, 1987.

[8] N. Menvielle, Y. Goussard, D. Orban, and G. Soulez, "Reduction of Beam-Hardening Artifacts in X-Ray CT," *27th Annual International Conference of the Engineering in Medicine and Biology Society, 2005.*, pp. 1865–1868, 2005.

[9] K. Jeong and J. Ra, "Reduction of Artifacts due to Multiple Metallic Objects in Computed Tomography," in *Medical Imaging 2009: Physics of Medical Imaging* (E. Samei and J. Hsieh, eds.), vol. 7258, p. 72583E, SPIE, 2009.

[10] G. Wang and M. Vannier, "Stair-step artifacts in three-dimensional helical CT: an experimental study," *Radiology*, vol. 191, no. 1, pp. 79–83, 1994.

[11] J. Barrett and N. Keat, "Artifacts in CT: Recognition and avoidance," *Radiographics*, vol. 24, no. 6, pp. 1679–1691, 2004.

[12] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, November 2004.

[13] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[14] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[15] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, 1997.

[16] C. Harris and M. Stephens, "A combined corner and edge detector.," in *Fourth Alvey Vision Conference*, pp. 147–151, 1988.

[17] D. Lowe, "Object recognition from local scale-invariant features," in *The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999.*, vol. 2, pp. 1150–1157, 1999.

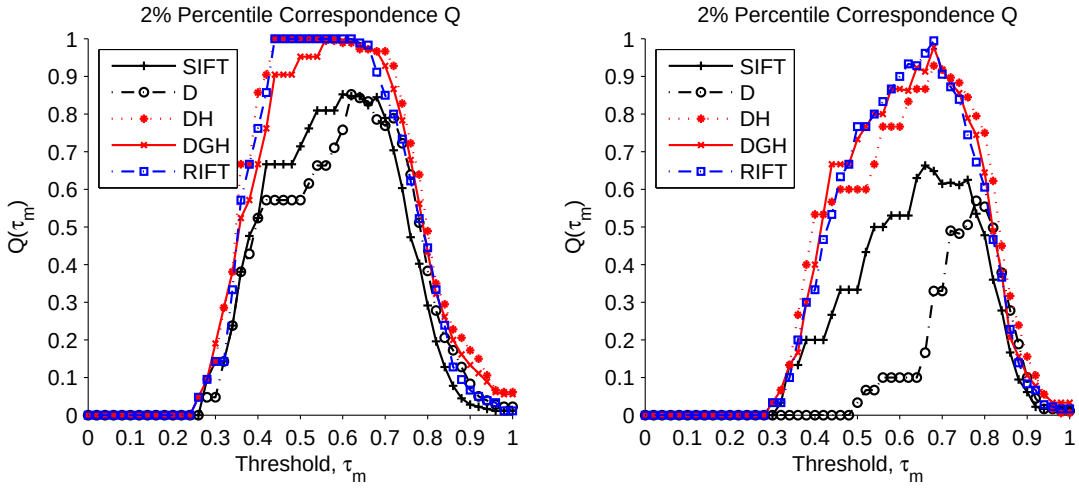
[18] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*, pp. 357–360, ACM Press New York, NY, USA, 2007.

[19] W. Cheung and G. Hamarneh, "N-Sift: N-Dimensional Scale Invariant Feature Transform For Matching Medical Images," *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2007*, pp. 720–723, April 2007.

[20] D. Ni, Y. Chui, Y. Qu, X. Yang, J. Qin, T. Wong, S. Ho, and P. Heng, "Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT," *Computerized Medical Imaging and Graphics*, vol. 33, no. 7, pp. 559–566, 2009.

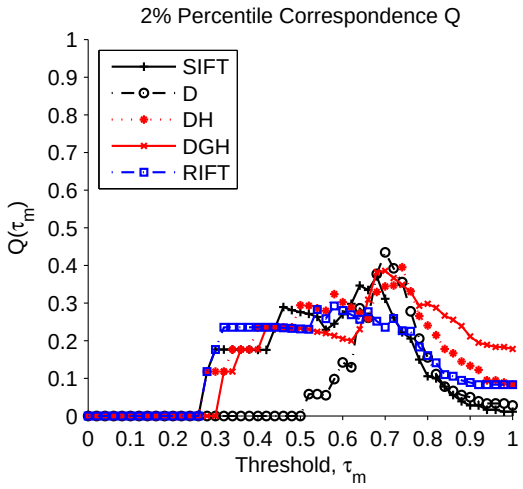
[21] S. Allaire, J. Kim, S. Breen, D. Jaffray, and V. Pekar, "Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008.*, pp. 1–8, June 2008.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "A Sparse Texture Representation

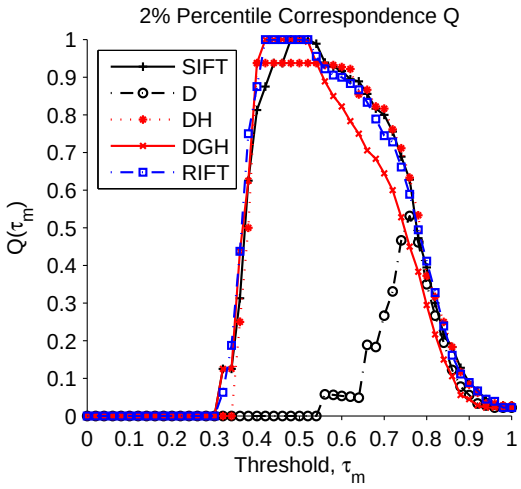


(a) Revolver

(b) Pistol (second reference)



(c) Apple iPod



(d) Binoculars

Figure 22: Threshold Quality for percentile $(p = 2\%)$ correspondence set

Using Local Affine Regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1265–1278, 2005.

[23] K. Mikolajczyk, B. Leibe, and B. Schiele, "Local features for object class recognition," in *Tenth IEEE International Conference on Computer Vision, 2005.*, vol. 2, pp. 1792–1799, 2005.

[24] C. Belcher and Y. Du, "Region-based SIFT approach to iris recognition," *Optics and Lasers in Engineering*, vol. 47, no. 1, pp. 139 – 147, 2009.

[25] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B. Lu, "Person-specific sift features for face recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.*, vol. 2, pp. 593–596, April 2007.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893 vol. 1, June 2005.

[27] M. Fischler and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[28] K. Arun, T. Huang, and S. Blostein, "Least-Squares Fitting of Two 3-D Point Sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 698–700, Sept. 1987.

[29] T. Fawcett, "An Introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[30] M. Schuckers, "Receiver Operating Characteristic and Equal Error Rate," in *Computational Methods in Biometric Authentication*, Information Science and Statistics, ch. 5, pp. 155–204, Springer London, 2010.

2013-02-16

A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery

Flitton, Greg

Elsevier

Flitton GT, Breckon TP, Megherbi Baouallagui N. (2013) A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery. Pattern Recognition, Volume 46, Issue 9, September 2013, pp. 2420-2436

https://doi.org/10.1016/j.patcog.2013.02.008